用于实时目标检测的高速可编程视觉芯片

李鸿龙[1,2], 杨　杰[1,2,3], 张忠星[1,2], 罗　迁[1,2], 于双铭[1,2], 刘力源[1,2], 吴南健[1,2,4]

(1. 中国科学院半导体研究所 超晶格国家重点实验室，北京 100083；
2. 中国科学院大学 材料与光电研究中心，北京 100049；
3. 西湖大学 工学院，浙江 杭州 310000；
4. 中国科学院脑科学与智能技术卓越创新中心，北京 100083)

摘　要：视觉芯片是一种高速、低功耗的智能视觉处理系统芯片，在生产生活中有广阔的应用前景。文中提出了一种新型的可编程视觉芯片架构，该架构的设计考虑了传统计算机视觉算法和卷积神经网络的运算特点，使其能够同时高效地支持这两类算法。该视觉芯片集成了可编程的多层次并行处理阵列、高速数据传输通路和系统控制模块，并采用 65 nm 标准 CMOS 工艺制程流片。测试结果表明：视觉芯片在 200 MHz 系统时钟下达到 413GOPS 的峰值运算性能，能够高效地完成包括完成人脸识别、目标检测等多种计算机视觉和人工智能算法。该视觉芯片在可编程度、运算性能以及能耗效率等方面都大大超越了其他视觉芯片。

关键词：视觉芯片；　目标检测；　卷积神经网络；　可编程阵列

中图分类号：TN492　　**文献标志码：**A　　**DOI：**10.3788/IRLA20190553

A high speed programmable vision chip for real-time object detection

Li Honglong[1,2], Yang Jie[1,2,3], Zhang Zhongxing[1,2], Luo Qian[1,2], Yu Shuangming[1,2], Liu Liyuan[1,2], Wu Nanjian[1,2,4]

(1. State Key Laboratory of Superlattices and Microstructures, Institute of Semiconductors,
Chinese Academy of Sciences, Beijing 100083, China;
2. Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China;
3. School of Engineering, Westlake University, Hangzhou 310000, China;
4. Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100083, China)

Abstract: Vision chip is a high-speed, low-power intelligent vision processing system on a chip (SoC), it can be applied to many different fields of production and daily life. A new programmable vision chip architecture was proposed which took into account the computational features of traditional computer vision algorithms and convolutional neural networks, it enabled the architecture to support the two types of algorithms simultaneously. The vision chip integrated multi-level parallel programmable processing array, high-speed data transmission path and system control module, it was fabricated in a 65 nm standard CMOS process. The experimental results show that the vision chip achieves 413GOPS peak performance at 200 MHz system clock and can perform various computer vision and artificial intelligence algorithms including face recognition and target detection efficiently.

收稿日期：2020−02−05；　**修订日期：**2020−03−25

基金项目：国家自然科学基金 (61434004, 61234003, 61504141, 61704167)；北京市重点研发计划 (Z181100008918009)；中国科学院青年创新促进会计划 (2016107)；中国科学院战略性先导科技专项 (XDB32050202)

作者简介：李鸿龙 (1990−)，男，博士生，主要从事高速视觉片上系统芯片方面的研究。Email：lihonglong@semi.ac.cn

导师简介：吴南健 (1961−)，男，研究员，博士生导师，博士，主要从事高速图像传感器、高速视觉片上系统芯片设计和图像并行处理片上系统芯片等方面的研究。Email：nanjian@red.semi.ac.cn

The proposed work shows higher system performance in terms of programmability, performance and energy efficiency when compared with other state-of-the-art vision chips.

Key words: vision chip; object detection; convolutional neural networks; programmable processing array

0 引 言

计算机视觉已经广泛应用于机器人导航、工业产品检测、自动驾驶、安防视频监控等诸多领域[1]。快速准确地实现图像分类、目标识别和追踪是计算机视觉领域的重要研究方向和热点，也是很多应用的关键所在。视觉芯片就是一种能够在多种场景下完成智能识别、追踪的高性能、低功耗、低延时的新型片上系统[2]。视觉芯片能够模仿人类视觉系统，集成了图像传感和处理的功能，相比于传统的基于通用 CPU、GPU 或者基于云端服务器的视觉处理系统，这种更靠近传感器的视觉处理特别适用于各种对体积、功耗和成本有严格限制的微小型嵌入式边缘型计算应用场景。

过去的几十年中，国内外报道了大量的视觉芯片研究工作。妙维等人报道了一种基于阵列型像素处理单元的视觉芯片，然而这种芯片只能完成简单的图像滤波、二值形态学处理任务[3]。张万成等人设计了集成多层次处理阵列的视觉芯片，能够完成低中高级不同复杂度的图像处理算法，但是识别速度无法满足实际应用的要求[4]。视觉芯片的架构设计跟随着算法和应用需求的发展而不断演变[3−7]。近年来，以卷积神经网络 (Convolutional Neural Networks, CNN) 为代表的深度学习算法在计算机视觉领域获得巨大的成功[8]，图像分类、检测的性能明显超越传统基于人工选取特征的算法。目前的视觉芯片设计已经很难胜任深度神经网络的处理，比如最近石匆等人提出的动态可重构视觉芯片，只能完成简单的 SOM 神经网络分类[5-6]，Yamazaki 等人设计的视觉芯片只能完成简单的卷积功能[7]。很显然，视觉芯片需要更加灵活的架构和电路设计以支持神经网络算法。深度神经网络算法存在着网络参数量大、运算量大和运算复杂等一系列的挑战，目前只有一些专门针对神经网络的加速器芯片可以完成神经网络所需的计算[9-10]。这类芯片只聚焦于对乘加运算的加速，在架构上并不适用于完成图像滤波、角点检测、直方图统计等重要的传统图像处理算法。然而，在各种视觉系统中，尤其是靠近图像传感器的应用中，经典的图像算法仍然是至关重要且必不可少的一环。

为了使视觉芯片具备神经网络计算能力，同时进一步提高其对各种传统视觉算法的支持，文中提出了一种新型的视觉芯片架构。该视觉芯片充分考虑了传统计算机视觉算法和卷积神经网络的计算特点，集成了灵活可编程的多层次并行处理阵列和高速数据传输通路以及系统控制模块，能够通过指令编程对各种计算资源进行灵活的重构和使用。视觉芯片采用 65 nm 标准 CMOS 工艺制造，经测试芯片在 200 MHz 的系统时钟下达到 413GOPS(Giga Operations Per Second) 的峰值性能，能够同时高效地支持传统的计算机视觉算法和卷积神经网络算法，实现实时的图像分类、目标检测、目标追踪等视觉处理任务。

1 视觉芯片的系统架构

1.1 算法的运算模式分析

为了能够很好地支持传统的计算机视觉和新兴的深度学习算法，文中首先详细分析了它们的运算特点，总结了实现这两类算法所需要的运算模式[5, 11]。

如图 1(a) 所示，传统的计算机视觉算法可以分解为三个运算层次。

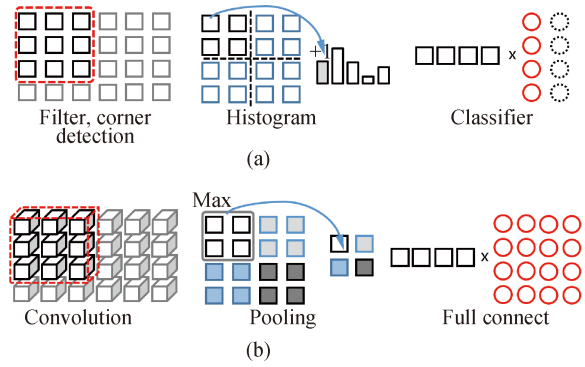


图 1 运算模式分析

Fig.1 Operation mode analysis

(1) 低层次处理—对图像进行滤波、边缘、纹理、

角点检测等处理。典型的滤波计算如下：

$$y(i,j)=\sum_{n}\sum_{m}I(i+m,j+n)*w(m,n)+b \tag{1}$$

式中：I 是输入图像；w, b 分别是是滤波核的系数和偏置；m, n 为对应的下标。每个像素和它周围 3×3 或者 5×5 邻域内的像素进行算术或者逻辑运算，每个点都可以独立计算，具有计算量大并行度高的特点。

(2) 中层次处理—对图像某一块区域进行信息统计，比如直方图、积分图统计。通常是把低层次处理后的图像分成若干区域进行，以 8×8 或者 16×16 大小的图像块为单位分别统计直方图之后得到特征向量。中层次处理的所需的运算较低层次处理更为复杂，但是并行度次之。

(3) 高层次处理—对特征向量进行分类，涉及到图像或者目标的整体信息，使用分类器判决图像是否属于要检测的目标类别。高层次处理运算最复杂但是并行度最低。一个典型的对输入向量x进行二分类的支持向量机运算如下：

$$y=\sum_{i}a_i y_i K(x_i,x)+b \tag{2}$$

式中：x_i是分类器的支持向量；a_i和y_i是支持向量的系数和标签，都由训练得到。对于最常用的线性核函数 $K(x_i,x)=x_i^{\mathrm{T}}x$，公式 (2) 可以写成：

$$y=w^{\mathrm{T}}x+b \tag{3}$$

式中：$w=\sum a_i y_i x_i$，主要的运算是向量内积。如果是多分类任务则会有多个分类器联合使用。

图 1(b) 给出了卷积神经网络算法的运算模式。卷积神经网络 (CNN) 主要由卷积层 (Convolution Layer, Conv)、激活层 (Activation Layer, Act)、池化层 (Pooling Layer) 和全连接层 (Fully Connected Layer, FC) 构成。通过反复使用卷积、激活、池化层来逐层进行信息的提取和抽象，最后用全连接层来进行分类。卷积层运算公式为：

$$y(i,j)=\sum_{c}\sum_{n}\sum_{m}I(i+m,j+n,c)*w(m,n,c)+b \tag{4}$$

式中：c 为图像通道 (channel)，除了在通道方向上进行累加外，卷积的运算模式和传统算法中的低层次处理滤波是一样的，可以逐个通道进行二维卷积公式 (1) 再相加即可得到公式 (4)。

激活层一般用在卷积层后面，主要是对卷积结果进行非线性处理，最常见的 ReLU 函数，只需要判断卷积结果的正负，更复杂的 Sigmoid 或者 tan$h$ 函数可以使用分段线性函数来逼近模拟。池化层是把图像划分为若干区域，计算每个区域中的平均值或者最大值，与中层次处理类似，都是以图像块为单位进行信息统计。

全连接层是实现向量和矩阵相乘，它的输入 x 是一个 n 维向量，输出是一个 m 维向量，那么全连接层的系数 w 是一个 $m*n$ 的矩阵，偏置 b 是一个 m 维向量：

$$y=wx+b \tag{5}$$

其运算模式和传统算法中的分类器类似，可以分解成若干个向量内积：

$$y_i=w_i^{\mathrm{T}}x+b_i \quad i=0,1\cdots m-1 \tag{6}$$

通过对公式 (1)~(6) 的分析，可以看到深度神经网络与传统的计算机视觉算法存在很多相似的运算模式，文中提出的新型视觉芯片架构充分利用这一特点，通过编程实现运算功能的重构，使得视觉芯片不仅能够兼容传统的基于人工选取特征和分类器的计算机视觉算法，而且能够高效地支持卷积神经网络等深度学习算法。

1.2 视觉芯片整体架构

文中提出了一种新型的基于多层次并行的可编程视觉处理芯片架构。图 2 是视觉芯片的整体架构图，整个芯片主要由图像计算阵列、数据传输通路和总体控制系统三大部分构成。图像计算阵列由多层次的并行处理单元和相应的存储器以及片上互联网络 (Network on Chip, NoC) 组成，主要负责完成各种算法的运算。数据传输通路主要由高速 IO 接口、片上 IO 数据缓存 (Input/Output Buffer) 和高速直接存储器存取 (Direct Memory Access, DMA) 构成，能够完成高速的片上/片下数据传输。RISC32 微处理器和系统总线以及其他相关模块是整个芯片的控制系统，指挥数据通路和计算阵列协同工作。整个视觉芯片是一个功能完整的视觉处理系统，具有高速低功耗、灵活可编程的特点，适合应用于对功耗、尺寸、实时性有较高要求的边缘计算等场合。

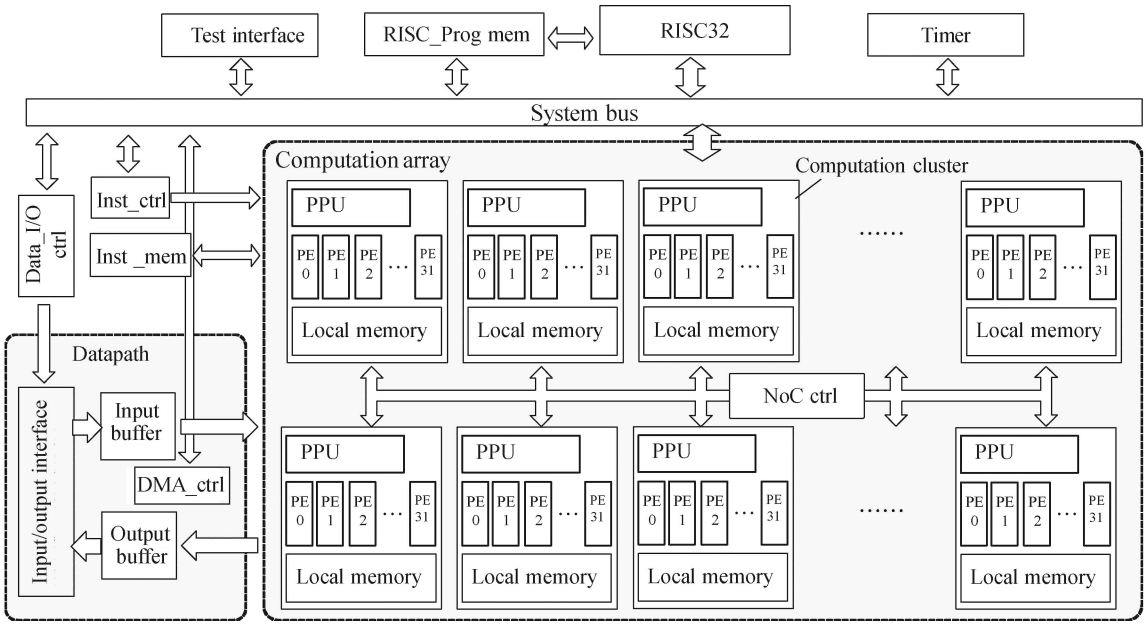


图 2 视觉芯片整体架构

Fig.2 Vision chip architecture

视觉处理系统的数据带宽要求非常高，由于很多应用场景都对图像传感器有特殊的要求，视觉芯片设计了高速的图像数据传输接口。芯片一共有 64 个数据输入和 64 个数据输出管脚，管脚的工作频率为 50 MHz。芯片其他部分的设计工作频率为 200 MHz，利用两块 IO 数据缓存作为异步 FIFO 进行数据跨时钟域传递。图像数据输入芯片之后先存放在 Input Buffer 中，然后用高速 DMA 模块把图像搬运到计算阵列中，开始进行图像数据的处理。计算阵列完成对图像的处理之后，DMA 模块再把处理结果从计算阵列搬到 Output Buffer，输出到片外。

计算阵列是视觉芯片进行图像数据运算的主要模块。如图 2 所示，整个计算阵列被分成了 16 组，每组被称为一个计算簇 (Computation Cluster)。每一个计算簇主要包含：一个块处理单元 (Patch Processing Unit, PPU)、32 个互相连接的 PE(Processing Element) 单元以及它们共享的本地存储器 (Local Memory)。图像数据进入处理阵列后会均匀分成 32×32 的小块分发到每个计算簇的本地存储器中，然后处理单元从本地存储器中读取数据进行运算。

计算簇的具体运算功能均由指令控制，视觉芯片的指令集采用了精简指令集的设计模式，并针对图像常用的运算进行了优化和改进。指令主要包含数据加载存储、数据运算和分支跳转三种功能，从处理并行度上分为标量和向量指令两种类型。块处理单元 (PPU) 是标量处理单元，功能比较灵活，能够获取 32×32 图像块中任意数据并进行计算。32 个 PE 构成向量处理单元，与一行图像 32 个像素一一对应，因此 PE 又被称为像素级处理器。32 个 PE 工作在单指令多数据 (Single Instruction Multiple Data, SIMD) 模式下，可以进行并行计算。

不同的计算簇也是工作在 SIMD 模式下，其中一个计算簇读取指令后分发给其他的计算簇。每个计算簇处理的是图像的不同部分或者卷积神经网络中不同的特征图映射 (feature map)，各个计算簇之间由 NoC 模块连接，可以广播或者交换数据，进行协同计算。

1.3 新型视觉芯片架构的特点

(1) 多层次并行的处理模式

计算阵列中的处理单元按照算法的需求来设计，计算簇中的向量处理器适合进行像素级并行处理，不同计算簇中的标量处理器在处理不同的图像块，构成了另一个多个层次的并行。整个视觉芯片上一共有 16 个块级处理器和 512 个像素级处理器。RISC32 微处理器除了控制系统之外还具有获取全局数据和进行复杂运算的能力，和计算簇中的块级处理器 (PPU)、像素级处理器 (PE) 一起构成了多层次并行的

处理阵列。不同层次的处理单元具有的计算能力和并行度是不同的，可以根据算法中不同层次的并行度需求进行调度。

(2) 动态可编程的运算功能

不管是传统的计算机视觉还是深度学习，算法都在不断地发展和变化，而在实际应用中也经常需要根据使用场景对准确率、速度的要求来选择使用的算法。整个计算阵列中的处理单元都是可编程的处理器核，其运算功能和连接状态都可以根据指令改变。通过基础指令的不同组合来实现不同的算法，可编程设计灵活地进行运算资源的动态配置，在不明显增加资源的情况下实现了对多种算法的支持。

(3) 高速灵活的数据传输

受限于片上存储的容量，通常输入图像的尺寸为 128×128 像素。如果要处理尺寸较大的图像或者参数较多的神经网络，需要把图像或者网络进行切分。比如在大图上使用滑动窗口逐块地处理或者是在计算卷积神经网络时逐层把参数传输进来计算。在处理第一帧数据的同时，芯片可以接收第二帧数据的输入。这样数据的传输和数据处理可以流水进行，隐藏了计算阵列等待数据的时间，图像处理结果的输出同理，整个传输和计算的调度由 RISC32 微处理器和系统总线进行控制。

2 电路设计

2.1 PPU 电路

图 3 展示了块处理单元 (PPU) 的具体电路。块处理单元是一个五级流水处理器，主要包含取指令、解码、执行、存储访问和写回五个部分。块处理单元具有比较灵活的处理功能，能够获取 32×32 本地存储的图像块中任意数据。除了串行处理的能力之外，PPU 还负责并将解码后的指令广播至向量处理器，控制指令循环跳转等功能，因此 PPU 是整个计算簇的控制核心。

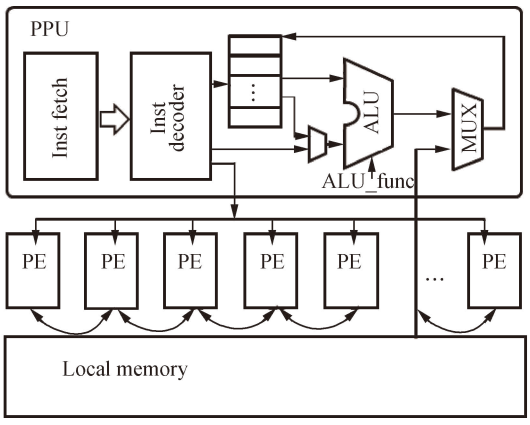


图 3 块处理单元电路图

Fig.3 Circuit of patch processing unit

2.2 PE 单元电路

图 4 是 PE 单元的电路图，每个 PE 具有一个 16 bit 的 ALU (arithmetic and logic unit) 和 16 个 16 bit 的寄存器组以及多个多路选择器。每个 PE 参与运算的数据可以来自于自身的寄存器中取，也可以来自左右邻近 PE 的寄存器组。ALU 的功能包括加、减、乘、乘累加、比较、移位、逻辑等，每周期能够进行一个 16 bit 或者两个 8 bit 的运算。与 PPU 单元相比，PE 单元没有取指令和指令解码的功能，它的运算的数据来源和运算功能均由 PPU 广播的指令控制，可以根据算法的需求实现不同的 PE 连接关系和运算功能。

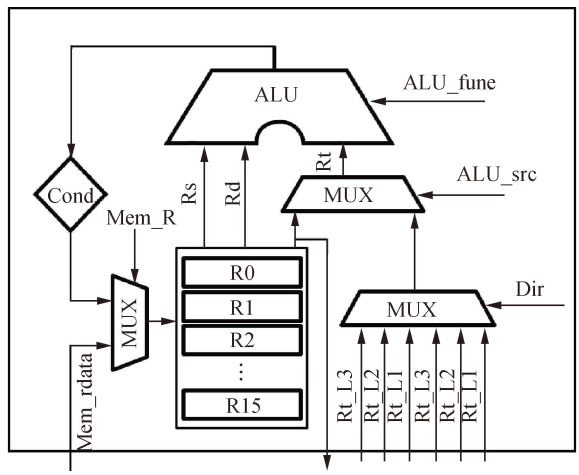


图 4 PE 单元电路图

Fig.4 PE circuit schematic

2.3 存储器电路

每一块本地存储器的容量为 16 kB，带宽为 256 bit，一个周期能读写 32 Byte 的数据。为了适应各种算法的需求，在存储器和 PE 单元之间有一个存储器接口 (memory interface) 模块来组织数据，实现各种复杂的读写功能。图 5 列举了一些读写的模式，如图 5(a) 中的正常模式常用于图像数据的读写，图 5(b) 中的两倍亚采样经常用于图像放缩和 CNN 中的池化层，图 5(c) 中的广播模式经常用于卷积系数的读取。

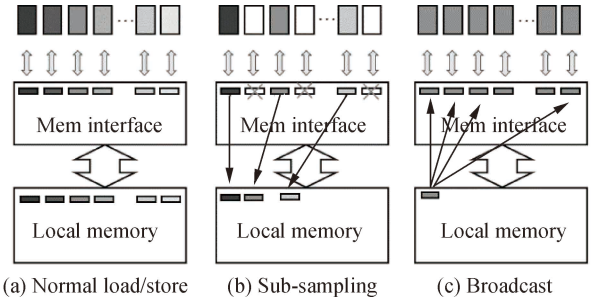


图 5 存储器的读写模式

Fig.5 Load/store pattern of memory

2.4 计算模式

(1) 像素级并行计算

像素级并行模式主要针对于图像的低层次处理，如滤波、角点检测、卷积层等。向量处理器横向排列，利用 PE 之间的相互连接关系，每个 ALU 的源操作数可以来自自身的寄存器组也可以来自邻近的 PE。这样的连接是特别为图像运算设计的。以进行 3×3 窗口的卷积运算为例，计算过程如图 6(a) 所示，只需要把三行像素加载到 PE 的寄存器组里，对于每个 PE 来说，寄存器组里只有纵向一列的三个像素，但是它能拿到两边邻近 PE 寄存器组中的数据，三个 PE 共同构成一个 3×3 的窗口，利用乘累加运算，九个周期就完成一行卷积的结果输出。

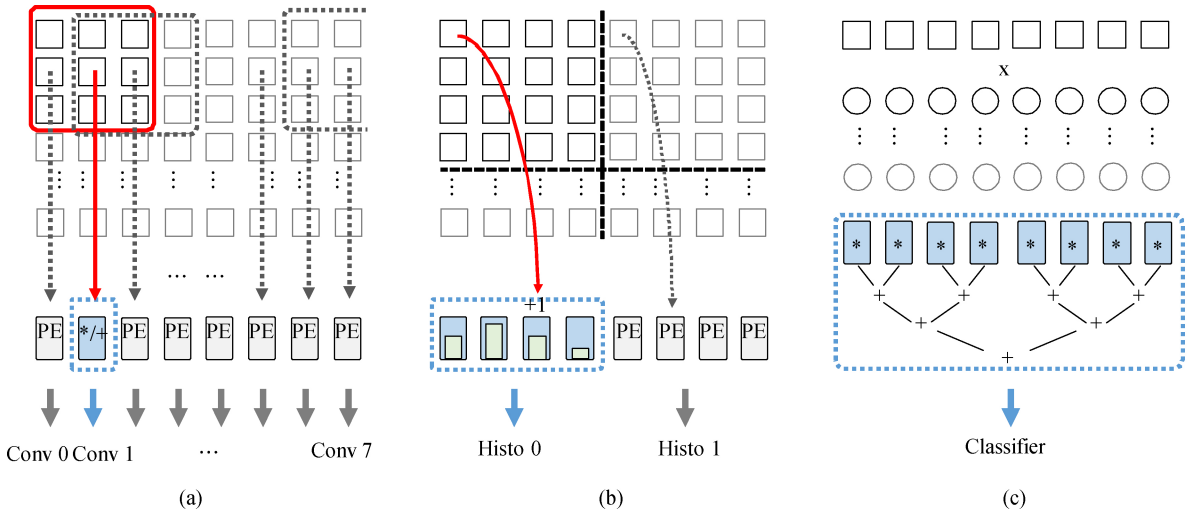


图 6 多层次并行运算

Fig.6 Multilevel parallel operation

每个 PE 都与左边三个、右边三个最邻近的 PE 连接，可以直接支持最大 7×7 的卷积窗口，阵列边缘的 PE 可以选择连接到 0 或者自身的寄存器组，实现卷积边缘填充 0 或者填充图像边缘的像素。把乘、加运算替换为减法、比较大小等其他基础运算，就可以实现梯度、纹理角点检测等其他算法。

可以看到，相邻的滤波 (卷积) 窗口之间存在大量的重叠区域，相互连接的 PE 阵列很好地利用了重叠区域数据的相关性，减少了数据的读取和重排。在像素级并行模式下每个 PE 都能进行一个卷积窗口的运算，并行度非常高。

(2) 块级并行计算

块级并行模式主要用于中层次的处理，比如直方图、池化层等，以图像块为单位，利用多个 PE 的组合来完成运算。图 6(b) 所示为直方图统计的计算过程，首先给每一个 PE 分配直方图的一个 bin，PPU 把像素广播到 PE 阵列，每一个 PE 分别对像素和分配的直方图 bin 值进行比较，如果 PE 的 bin 和像素值相等，则把计数器+1，其他的 PE 计数器则保持原值。

使用 PE 本身的寄存器组作为计数器，可以避免存储器的反复读写，使统计具有很高的效率。计算池化层的时候，每个 PE 先纵向找出一列里的最大值，然后再横向比较，得到图像块里的最大值，最后用亚采样存的方式输出池化结果。根据直方图 bin 的数目或者是池化区域的大小，把 PE 阵列划分为若干组，每组负责一个图像块的统计，可以达到块级并行的处理效果。

(3) 全局并行计算

全局并行模式主要用于分类器的向量内积计算

或者全连接层的向量矩阵相乘，计算同样可以利用 PE 间的连接关系，使所有的 PE 共同参与计算。向量内积计算过程如图 6(c) 所示，首先每个 PE 分别计算两个向量中对应元素的乘积，然后利用 PE 之间的连接把相邻两个 PE 的乘积相加，再利用亚采样存储的指令把间隔 PE 的和保留下来。反复进行这样两两合并的运算，对于 N 维的向量合并速度为 $\log_2 N$，即使对于几百维甚至上千维的特征向量，也只需要几十个时钟周期就能完成向量内积的计算。最后利用 RISC32 对分类器的计算结果进行汇总和比较，筛选最合适的分类结果。

视觉芯片从整体框架到具体电路的设计都遵循了软硬件协同的原则，不管是传统的基于手工选择特征的算法还是深度神经网络算法，它们所需的所有基本运算操作都可以通过控制 PE 之间的连接关系和 ALU 的运算功能来实现。整个芯片具有很好的可编程性，可以根据应用的需求灵活地选择算法。

3 芯片实现与测试结果

3.1 芯片实现和测试系统

可编程视觉芯片采用 1P9M 65 nm CMOS 工艺制成，芯片大小为 4 mm×6 mm，其显微照片如图 7 所示。芯片周围是电源/地线、数据传输和控制信号管脚，16 个计算簇对称地排放，中间则是数据通路以及控制系统。

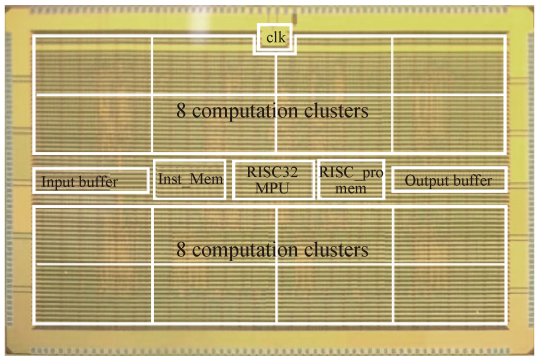


图 7 视觉芯片显微照片

Fig.7 Microphotograph of the vision chip

经测试，芯片在 1 V 标准供电电压下运行频率达到 200 MHz，在进行乘法器利用率最高的卷积神经网络计算时功耗仅 1.07 W。表 1 列出了该芯片的重要性能参数。当进行 16 bit 精度运算时，芯片峰值运算性能达到 208GOPS。采用 8 bit 精度运算时，芯片峰值运算性能则可到达 413GOPS。

表 1 视觉芯片参数

Tab.1 Vision chip parameters

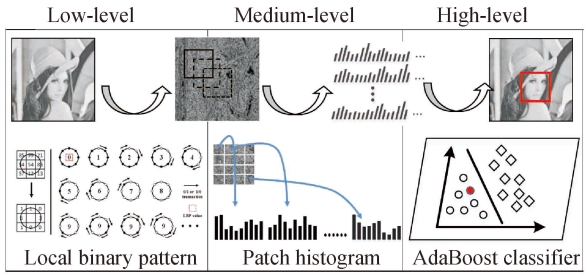
Technology	65 nm 1P9M
Size	4 mm×6 mm
Core frequency	200 MHz
Processing unit	1+16+512
Operand bit-width	2×8-bit fixed / 16-bit fixed
Peak performance	413GOPS@8b / 208GOPS@16b
Peak power	1.07 W
I/O bandwidth	3.2 Gbps (duplex)

3.2 算法测试

基于前面对算法模式的分析，该节采用一些应用算法对视觉芯片进行系统级的性能测试。为了体现视觉芯片的可编程性，分别选取了使用传统识别算法、深度学习算法和两者联合使用的应用对视觉芯片进行算法测试。

3.2.1 人脸检测

图 8 给出了基于 LBP(Local Binary Pattern) 特征和 Adaboost 分类器的人脸检测流程和测试结果。图 8(a) 展现了一个典型的包含特征生成、特征提取以及特征分类三个步骤的传统目标检测算法。首先在 3×3 的窗口内，每个中心像素与周围八个像素比较大小，生成 8 bit 的 LBP 值；然后按 16×16 的图像块统计 LBP 直方图，形成特征向量；最后使用 Adaboost 分类器判别是否为人脸。计算 LBP 值、统计直方图和分类器分别对应低、中、高层次的处理，在视觉芯片上均可以高效支持。目标搜索框的大小为 64×64，以步长为 16 像素在 128×128 的图像中移动进行检测，在判别过程中如果有多个重叠度大于 0.5 的搜索框都判定为人脸，则选取分类器响应最大的候选框作为检测结果。在检测的过程中 LBP 值和分块直方图都只需要计算一次，选择搜索框中的直方图作为特征向量进行分类即可。计算整张图 LBP 值和直方图分别需要大约 2 400 和 1 600 个周期。逐个搜索组合特征向量进行分类一共需要 6 000 个周期，再加上最后筛选结果并输出，计算一幅图像大约 12 000 个时钟周期，检测速度超过 16 000 fps。



(a)

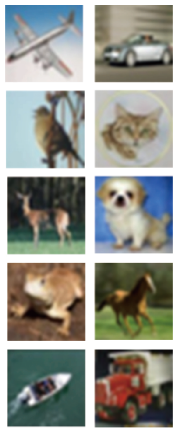
(b)

图 8 人脸识别算法和测试结果

Fig.8 Face detection algorithm and test result

3.2.2 CNN 图像分类

图 9 给出了基于 cifar10 数据集的 CNN 分类网络。网络采用 16 bit 定点参数，预先线下训练好参数并进行定点化，为了适应硬件实现的特点，激活函数采用简单的 ReLU 函数，只需要判断卷积结果的正负，池化层都采用两倍下采样，这样卷积神经网络中的每一种操作视觉芯片都能高效支持。在计算卷积的过程中，输入的图像尺寸是 32×32，对于不同的卷积核，每个计算簇可以计算一个 feature map，随着池化层的加入，尺寸缩小，每个计算簇可以同时计算两个或四个 feature map。经测试对于 32×32 的 RGB 图像，



Layer	Kernel size	Output size
Input		32×32×3
Conv 1	5×5×3×32	28×28×32
Max P1	2×2	14×14×32
Conv 2	3×3×32×32	12×12×32
Conv 3	3×3×32×64	10×10×64
Max P2	2×2	5×5×64
Conv 4	3×3×64×64	3×3×64
FC1	3×3×64×64	1×1×64
FC2	1×1×64×10	1×1×10

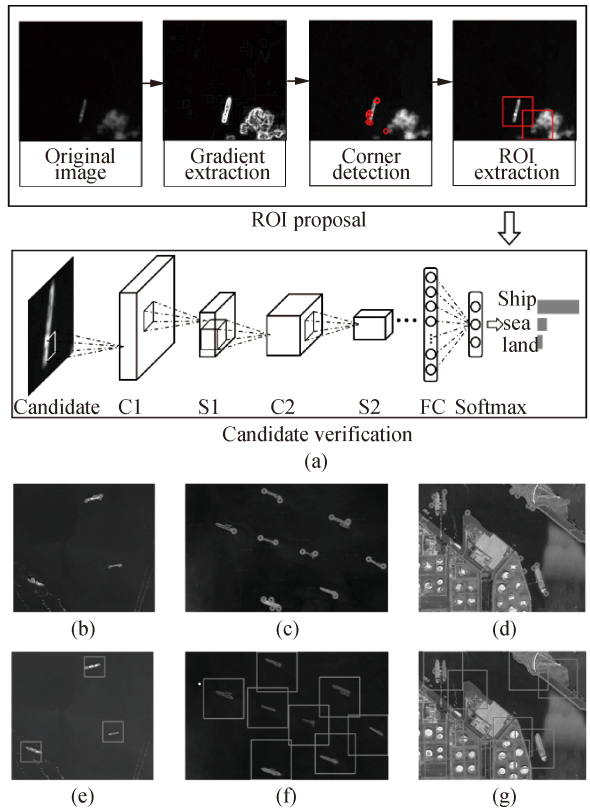
图 9 CNN 图像分类

Fig.9 CNN image classification

计算四层卷积和两层全连接的网络需要大约 26 000 个时钟周期，处理速度达到 7 000 fps。

3.2.3 遥感图像目标检测

在海洋遥感图像中视野非常宽广，大部分是海面和云层，目标的占比很小。采用传统的方法准确率比较低，如果全部采用卷积神经网络检测效率太低，速度太慢。文中设计了一种两级级联检测的方法，如图 10 所示，首先用传统的梯度和角点检测，设置合适的阈值筛掉大部分没有什么亮度变化的平静海面和大片云层，提取出可能存在船只的感兴趣区域 (region of interest, ROI)，然后对留下来的部分使用 CNN 进行精细分类，辨别是否是为船只。这样的混合算法速度比全图使用 CNN 搜索检测更快而准确率基本保持不变。图 10 展示了在遥感图像中进行船只检测的算法流程和测试结果。



(a)

(b) (c) (d)

(e) (f) (g)

图 10 遥感图像目标检测

Fig.10 Target detection in remote sensing image

3.3 性能分析与比较

测试结果表明，文中提出的视觉芯片确实能实现从图像预处理到特征提取再到特征分类的完整视觉处理流程，而且可以通过编程灵活实现不同的算法应

用,达到实时检测的系统性能。表 2 列出了一些相关的视觉芯片和神经网络加速器芯片的参数对比。

表 2 与其他相关工作的对比

Tab.2 Comparison with related work

References	This work	Ref [5]	Ref [12]	Ref [13]	Ref [14]
Technology/nm	65	180	130	65	65
Core area/mm^2	4×6	9.8×8.4(include sensor)	187	4×4	4×4
Frequency/MHz	200	50	80	200	250
On-chip SRAM	320 kB	32 kB	236 kB	192 kB	282 kB
Bit-width(fixed)	1/8/16	1~16	8	16	16
Traditional computer vision	Filter, corner detection Histogram SVM/Adaboost	Filter, corner detection Histogram	Filter(enhance, morphology)	N/A	N/A
Convolution neural network	Conv Pooling FC	SOM neural(FC)	Conv	Conv	Conv Pooling FC
Peak performance	413GOPS@8b 208GOPS@16b	14GOPS@8b	61GOPS@8b	67GOPS@16b	32GOPS@16b
Energy-efficiency (GOPS/W)	194 @16b	44 @8b(processor array)	85 @8b	166 @16b	208 @16b

通过表 2 可以看出,与近年来所报道的其他视觉芯片相比,文中所提出的视觉芯片架构在片上存储容量、算法支持和峰值运算能力上有明显优势。特别是在对算法多样性的支持方面,除了能够很好地保留以往视觉芯片支持的传统计算机视觉算法之外,还能高效地支持卷积神经网络等深度学习算法,即使与一些专用的神经网络加速器相比,能耗效率依然不落后。文中提出的视觉芯片在可编程性、运算性能以及功耗效率等方面都超过了其他视觉芯片。

4 结 论

文中提出了一种新型的高速可编程视觉芯片架构,它包括多层次并行的可编程处理阵列、高速数据传输通路和总体控制系统,不仅能够兼容传统的基于人工选取特征和分类器的目标检测算法,而且能够高效地支持卷积神经网络等深度学习算法。视觉芯片由 65 nm 标准 CMOS 工艺制成,在 200 MHz 系统时钟下达到 413GOPS 的运算性能和 194GOPS/W@16 bit 的能耗效率,在处理能力、可编程能力和能耗效率等方面明显优于其他类似的视觉芯片。测试结果表明:文中所设计的视觉芯片能够完成完整的图像分类、目标检测等视觉处理流程,实现了较高的系统性能。

参考文献:

[1] Sonka M, Václav Hlavác, Boyle R. Image Processing, Analysis and Machine Vision[M]. New York: Cengage Learning, 2014.

[2] Ishikawa M, Ogawa K, Komuro T, et al. A CMOS vision chip with SIMD processing element array for 1 ms image processing[C]//IEEE Int Solid-State Circuits Conf, 1999: 206-207.

[3] Miao W, Lin Q, Zhang W, et al. A programmable SIMD vision chip for real-time vision applications [J]. ***IEEE Journal of Solid-State Circuits***, 2008, 43(6): 1470−1479.

[4] Zhang W, Fu Q, Wu N. A programmable vision chip based on multiple levels of parallel processors [J]. ***IEEE Journal of Solid-State Circuits***, 2011, 46(9): 2132−2147.

[5] Shi C, Yang J, Han Y, et al. A 1000 fps vision chip based on a dynamically reconfigurable hybrid architecture comprising a PE array and self-organizing map neural network[C]//2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014: 128-129.

[6] Yamazaki T, Katayama H, Uehara S, et al. 4.9 A 1 ms high-speed vision chip with 3D-stacked 140GOPS column-parallel PEs for spatio-temporal image processing[C]//2017 IEEE International Solid- State Circuits Conference (ISSCC), 2017: 82-83.

[7] Yang J, Yang Y, Chen Z, et al. A heterogeneous parallel processor for high-speed vision chip [J]. ***IEEE Transactions on Circuits and Systems for Video Technology***, 2018, 28(3): 746−758.

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in

Neural Information Processing Systems. 2012: 1097-1105.

[9] Sze V, Chen Y, Yang T, et al. Efficient processing of deep neural networks: A tutorial and survey [J]. ***Proceedings of the IEEE***, 2017, 105(12): 2295−2329.

[10] Chen T, Du Z, Sun N, et al. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning [J]. ***ACM Sigplan Notices***, 2014, 49(4): 269−284.

[11] Liu D, Chen T, Liu S, et al. Pudiannao: A polyvalent machine learning accelerator[C]//ACM SIGARCH Computer Architecture News. ACM, 2015, 43(1): 369-381.

[12] Millet L, Chevobbe S, Andriamisaina C, et al. A 5500FPS 85GOPS/W 3D stacked BSI vision chip based on parallel in-focal-plane acquisition and processing[C]//2018 IEEE Symposium on VLSI Circuits. IEEE, 2018: 245-246.

[13] Chen Y, Krishna T, Emer J, et al. 14.5 Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks[C]//2016 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2016: 262-263.

[14] Jo J, Cha S, Rho D, et al. DSIP: A scalable inference accelerator for convolutional neural networks [J]. ***IEEE Journal of Solid-State Circuits***, 2017, 53(2): 605−618.